Counterfactual generating Variational Autoencoder for Anomaly Detection

Renate Ernst[1,2]

[1]Fraunhofer Institute for Industrial Mathematics (ITWM), Fraunhofer-Platz 1, 67663, Kaiserslautern, Germany
[2]University of Kaiserslautern-Landau (RPTU), Gottlieb-Daimler-Straße, 67663, Kaiserslautern, Germany

Abstract

Machine learning applications in fields such as financial accounting or the healthcare industry have to meet high transparency requirements for user acceptance and to meet the growing number of regulatory standards. Counterfactual explanations as a rather easy to interpret concept of local explanations combined with the generative power of Variational Autoencoder (VAE) and their ability to learn distributions of latent representations can offer information to fulfill the needs of machine learning experts and non-expert users at the same time. Most current studies leveraging the power of deep generative models for counterfactual generation focus on vision data. We focus on anomaly detection applications on real world tabular data in the two high-risk fields of financial accounting and healthcare. We give an overview on constructions of counterfactual explanations and a categorization of current approaches to produce counterfactual explanations. We are investigating supervised extensions of the VAE for simultaneous classification and counterfactual generation. Therefor we explore the connection between different approaches of probabilistic modelling and separability properties in latent space. We discuss their applicability to anomaly detection and evaluation criteria.

Keywords

explainable AI, variational autoencoder, counterfactual explanation, anomaly detection

1. Introduction

Generative neural networks as a special form of deep learning have changed the way of data driven application in many fields, such as computer vision, robotics, and natural language processing. Specifically medical and financial industries ML-based approaches for decision support have to meet high requirements of transparency. Explainability methods are constructed to help the user of ml-methods to interpret the results of black-box models. In this PhD project we focus on counterfactual explanations as an local explanability approach and investigate, how this concept can be used to answer the users question: "What Should I have done differently to change the outcome of the model prediction?" We investigate how counterfactual explanations can be integrated into the VAE method to generate exogenous counterfactuals in the context of anomaly detection.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta
renate.ernst@itwm.fraunhofer.de (R. Ernst)
https://www.itwm.fraunhofer.de/de/abteilungen/fm/mitarbeiter/renate-ernst.html (R. Ernst)
0009-0006-7484-4638 (R. Ernst)

2. Related work

The VAE was introduced by [1]. It is based on a directed causal model using Bayes statistic to learn the parameters of the distribution of a latent variable. It can be used to detect anomalies in an unsupervised setting, by training the model on normal data and using the reconstruction error as anomaly score [2]. The concept of counterfactual explanations was formulated as an mathematical optimization problem for ML-models by [3]. There is no consensus in the scientific literature on the taxonomy of explainability. Based on [4], we will interpret counterfactuals as local instance explanations as an post hoc approach and a way of improving interpretability of an ML-model as an ante hoc approach. The VAE as post hoc approach is visualized in figure 1 by the green lines, and as an ante hoc approach by the pink line.

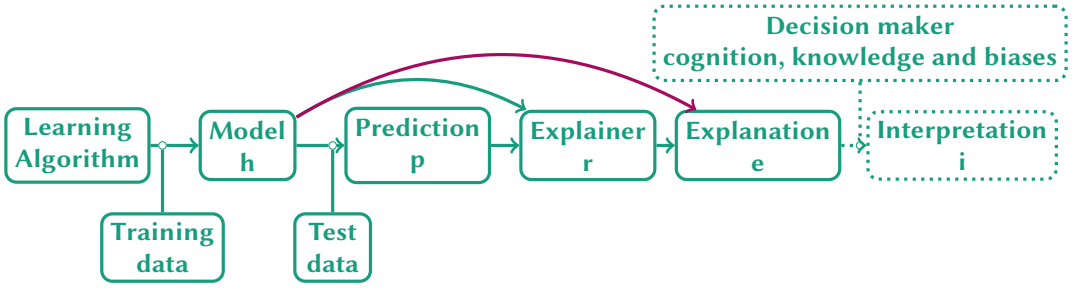


Figure 1: Flow chart from learning algorithm to human interpretation with post hoc explainer.

An explanation is called interpretable, when it describes the internals of a system in a way that is understandable to humans. The success of this goal is tied to the cognition, knowledge, and biases of a decision maker. For an explanation to be interpretable, it must give descriptions that are simple enough for a decision maker to understand using a vocabulary that is meaningful to the them. Note, that we will not measure the interpretability of a model. But we measure, how well the construction goals of the counterfactual are met by the approach used.

An overview over current approaches for calculating counterfactual explanations is given in [5]. The approaches can be categorized into model agnostic and model specific approaches. We will look at concepts specific to VAE.

3. Research goals

This project aims to achieve the following goals:

1. get an overview on counterfactual construction objectives (CE-objectives) and categorize, how they can be integrated into a VAE training procedure,
2. develop metrics to evaluate, how good the CE-objectives can be achieved,
3. develop the best way to integrated the CE-objectives into the VAE-objective specifically for anomaly detection and
4. apply our method(s) on real world anomaly detection use cases from the healthcare and financial sector.

4. Methods

4.1. Counterfactual Explanations

Counterfactual explanations (CE) have a long history in philosophy, but [3] first mathematically formulated the counterfactual explanation as solution of an optimization problem for machine learning.

Definition 4.1 (Counterfactual explanation). *Given a classifier b that outputs the decision $y = b(x)$ for an instance x, a counterfactual explanation consists of an instance x' such that the decision for b on x is different from y, i.e., $b(x) \neq y$ and such that the difference between x and x' is minimal.*

Since the first formulation, the requirements on this concept have evolved to meet practical considerations. In [5] these CE-objectives are formulated as:

1. **Validity:** A counterfactual x' should actually changes the classification outcome.
2. **Proximity:** Given a distance function d in the domain of x, the distance between x and x' should be as small as possible.
3. **Minimality:** There should not be any other valid counterfactual example x'' such that the number of different attribute value pairs between x and x' is higher than the number of different attribute value pairs between x and x''.
4. **Plausibility:** The counterfactual x' should not be labeled as an outlier with respect to the instances in X.
5. **Diversity:** Let $C = \{x'_1, \dots x'_k\}$ be a set of k (valid) CE for the instance x. The CE C should be formed by diverse CE, i.e., while every CE $x'_i \in C$ should be minimal and close to x, the difference among all the CE in C should be maximized.
6. **Actionability:** A CE x' is actionable if all the differences between x and x' refers only to actionable (mutable) features. This requirement links the concept of counterfactual explanation to the concept of algorithmic recourse.
7. **Causality:** Let G be a directed acyclic graph (DAG) where every node models a feature and there is a directed edge from i to j if i contributes in causing j. The DAG G describes the known causalities among features. Thus, given a DAG G, a counterfactual x' respects the causalities in G iif $\forall x'_i = (a_i, v_i) \in \delta_{x,x'}$ such that the node i in G has at least an incoming/outcoming edge, the value v_i maintains any known causal relation between i and the values $v_{j_1}, \dots, v_{j_m}$, where the features $j_1, \dots, j_m$ identifies the nodes connected with i in G.

4.2. Variational autoencoder

The VAE was introduced in [1] as following. Consider dataset $X = \{x^{(i)}\}_{i=1}^{N}$ consisting of N i.i.d. samples of some continuous or discrete variable x. We assume, the data is generated by some random process, involving an unobserved continuous random variable z. The process consists of two steps:

- a value $z^{(i)}$ is generated from some prior distribution $p_{\theta^*}(z)$.

- a value $x^{(i)}$ is generated from some conditional distribution $p_{\theta^*}(x|z)$.

We assume the prior $p_{\theta^*}(z)$ and likelihood $p_{\theta^*}(x|z)$ come from parametric families of distributions $p_\theta(z)$ and $p_\theta(x|z)$, and that their probability density functions (PDF(s)) are differentiable almost everywhere w.r.t. both θ and z. The true parameters θ^* as well as the values of the latent variables $z^{(i)}$ are unknown. Where the integral of the marginal likelihood $p_\theta(x) = \int p_\theta(z)p_\theta(x|z)dz$ is intractable, the true posterior density $p_\theta(z|x) = p_\theta(x|z)p_\theta(z)/p_\theta(x)$ is intractable. Let us introduce a surrogate $q_\phi(z|x) \approx p_\theta(z|x)$ to approximate the intractable true posterior. We refer to $q_\phi(z|x)$ as probabilistic encoder and $p_\theta(x|z)$ as probabilistic decoder.

The VAE-objective aims to maximize the so called evidence lower bound (ELBO).

$$\mathcal{L}(\phi, \theta; x) = -D_{KL}(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right], \tag{1}$$

where D_{KL} is the Kullback-Leibler-divergence (KLD).

By using a neural network architecture for encoding and decoding and neural network optimization (e.g. adam) [1] use the power of neural network optimization to autoencode variational bayes.

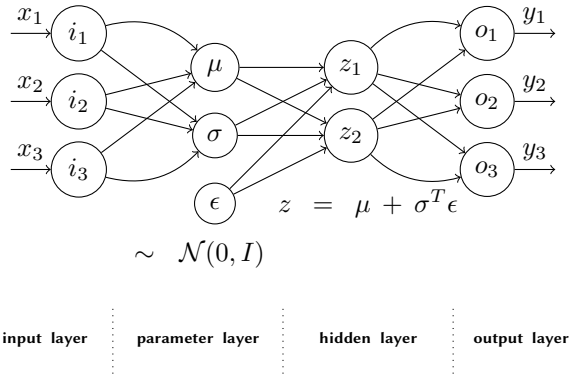


Figure 2: Illustration of a minimal example VAE architecture

The VAE can be used in an unsupervised setting to perform anomaly detection (see [2]), by training the model on normal data. The ability of the model to reconstruct the test data measured by the reconstruction error can be used as anomaly score.

4.3. Integration of CE-objectives into the VAE training

In the current literature we can see modifications to the VAE architecture, to produce CE. There are modifications used to increase the interpretability of the network architecture. For example the use of invertable mappings form feature space to latent space (see [6] or [7]) or to encode a hierarchical latent sequence (z_k) to capture more complex causal structures in latent space (see [8]). There are modifications to the VAE-objective to include CE-objectives. This can be included by regularization (see [7] or by conditioning the prior distribution to control the CE candidate generation(see [9]). If the model is not explicitly trained to generate counterfactuals, the training objective aims to separate the different classes and use some perturbation (see [10]),

projection (see [6]), interpolation (see [8]) or sampling technique (see [11]) to generate CEs. Due to reference space limitations, the overview of current literature is not comprehensive, but all current methods can be categorised into one of the following tree categories:

- Include the CE-objectives in the VAE-objective and train the model to produce a CE.
- Train the VAE to separate the classes. Use perturbation or projection technique to produce valid candidate(s). Select the candidate optimal under the CE-objectives.
- Train different models on partitioned data. Use interpolation or sampling technique to produce valid candidates. Select the candidate optimal under the CE-objectives.

Since the methods are designed for different data types, use data specific model modifications and try to meet different CE-objectives, a comprehensive comparison using the literature so far is not possible. None of the methods is designed or evaluated for anomaly detection use-cases with rare anomaly data.

5. Preliminary results

After reviewing and grouping current approaches on counterfactual generation in VAE, we develop ideas on how to integrate the CE-objectives specifically for the anomaly detection use-case. Currently we are investigating the complementary supervised VAE approach introduced in [12]. We want to investigate it's ability to separate anomalies from normal data and using it, to generate CE candidates. The concept in [12] is to train the VAE with normal prior and a loss function with so called standard KLD on a dataset with normal data. For training on seen anomaly data a complementary prior and corresponding KLD is chosen. The complementary set VAE approach follows the assumptions, that anomalies are regarded as the complementary set of the normal set and the normal region and the anomalous region are both mutually exclusive and collectively exhaustive. With defining $p_n(z)$ to be the PDF for normal data [12] construct $p_a(z)$ to be PDF of the anomalous data. It is constructed to satisfy the relationship

$$p_a(z) = \frac{1}{k'}(\max_{z'} p_n(z') - p_n(z)) \tag{2}$$

where k' is a norming constant such that $p_a(\cdot)$ is PDF. This construction satisfies the property of the complementary set, but k' is infinity because the mass explodes. To ensure $p_a(z)$ is a PDF, we multiply $p_w(z)$ that is wide enough for each dimension. Then the density function is

$$p_a(z) = \frac{1}{k} \underbrace{p_w(z)(\max_{z'} p_n(z') - p_n(z))}_{=:p_a^*(z)} \tag{3}$$

where k is a finite normalizing constant

$$k = \int_{-\infty}^{\infty} p_a^*(z)dz. \tag{4}$$

Using this as a prior, [12] expand the conventional unsupervised VAE into a supervised one to distinguish anomalies in the latent space. We choose the Standard Gaussian distribution

as a prior for the representation of normal samples $z_n \sim \mathcal{N}(z_n; 0, 1)$ with PDF $p_n(z_n; 0, 1)$. We construct the one-dimensional PDF $p_a(z_a)$ of the representation for anormal samples z_a using the prior for the normal data representation $p_n(z_a; 0, 1)$, the bounding Gaussian density function $p_w(z_a; 0, s^2)$ and equation (3)

$$p_a(z_a; s^2) = \frac{1}{k} p_w(z_a; 0, s^2) \cdot (\max_{z'_a} p_n(z_a; 0, 1) - p_n(z_a; 0, 1)) \tag{5}$$

where the constants in this equation are described as

$$\max_{z'_a} p_n(z_a; 0, 1) = \frac{1}{\sqrt{2\pi}} =: a \tag{6}$$

and

$$k = \int_{-\infty}^{\infty} p_w(z_a; 0, s^2) \cdot (a - p_n(z_a; 0, 1)) dz_a = a \left(1 - \sqrt{\frac{1}{s^2 + 1}}\right). \tag{7}$$

The parameter s^2 determines the width of the distribution. The multi-dimensional version is derived as a product of each dimension composed of the one-dimensional version. The

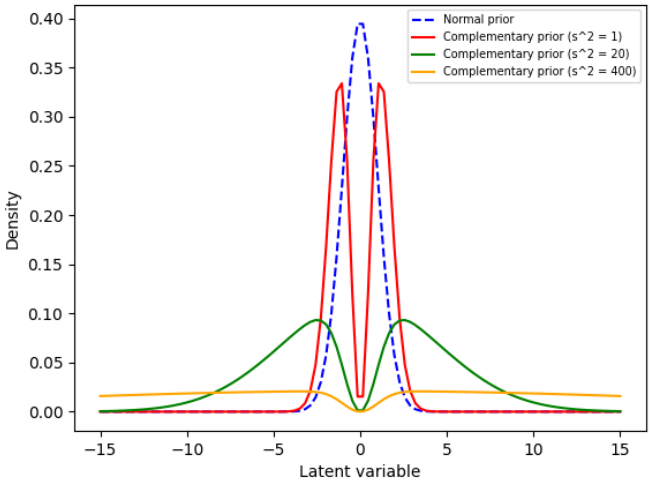
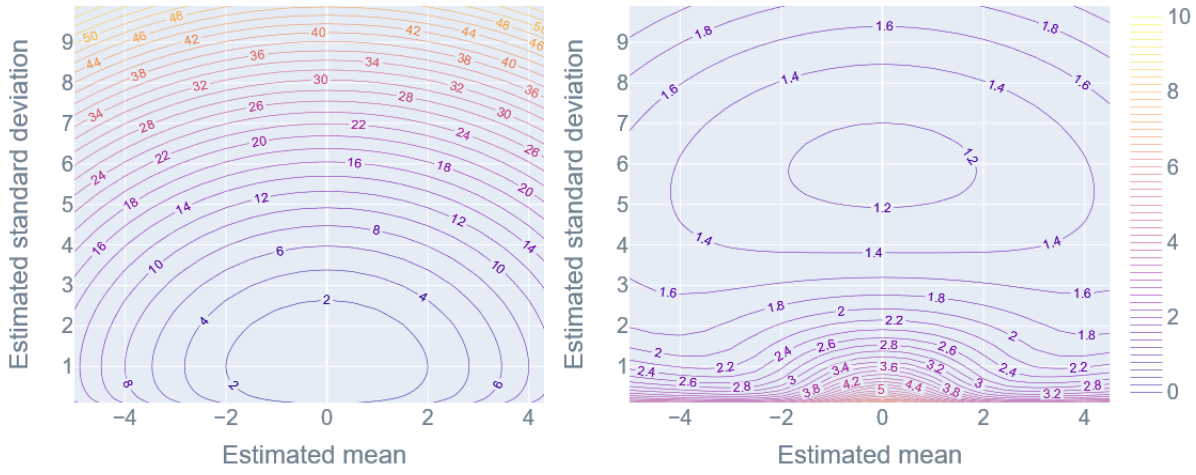


Figure 3: Visualisation of complementary prior for different s^2 in one dimension.

complementary KLD can be approximately calculated as

$$KL(q(z|x;\phi)||p_a(z)) \simeq \sqrt{\frac{2\pi}{\sigma^2 + 1}} \exp\left(\frac{-\mu^2}{2(\sigma^2 + 1)}\right) + \frac{\mu^2 + \sigma^2}{2s^2} - \log\sigma + \log s + \log(\sqrt{s^2 + 1} - 1) - \frac{\log(s^2 + 1)}{2} + \frac{\log(2\pi) - 1}{2} \tag{8}$$



(a) Standard KLD

(b) Complementary KLD for s = 5

Figure 4: Visualization of the KLD for the normal prior and anomaly prior

This VAE is trained alternating on a batch of normal data with standard prior and KLD and a batch of anormal data with complementary prior and complementary KLD. It is based on the idea, that both models share parameters. We want to investigate this training setup, because it can be used to detect also unseen anomalies. We have implemented the training procedure and want test it on synthetic data with given distributions, to validate this training setup.

6. Next research steps and expected final contribution

We aim to give an overview and discuss different evaluation metrics for current approaches and for our model extensions. In next research steps we evaluate the training procedure for complementary set VAE, produce and evaluate CE-candidates and develop and evaluate different approaches to integrate CE-objectives in the VAE-objective. Specifically we aim to investigate modifications in conditioning the prior distribution and the causal model of the VAE, investigate modifications in optimization, e.g multi-criteria optimization, investigate different training settings and develop and discuss an evaluation scheme for synthetically generated data and for real world use case data.

References

[1] D. P. Kingma, M. Welling, Auto-encoding variational bayes, 2022. URL: https://arxiv.org/abs/1312.6114. arXiv:1312.6114.

[2] Jinwon An, Sungzoon Cho, Variational autoencoder based anomaly detection using reconstruction probability, 2015. URL: http://dm.snu.ac.kr/static/docs/tr/snudm-tr-2015-03.pdf.

[3] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, SSRN Electronic Journal (2017). doi:10.2139/ssrn.3063289.

[4] N. Burkart, M. F. Huber, A survey on the explainability of supervised machine learning, Journal of Artificial Intelligence Research 70 (2021) 245–317. URL: https://www.jair.org/index.php/jair/article/view/12228. doi:10.1613/jair.1.12228.

[5] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022) 1–55. doi:10.1007/s10618-022-00831-6.

[6] W. Zhang, B. Barr, J. Paisley, An interpretable deep classifier for counterfactual generation, in: D. Magazzeni, S. Kumar, R. Savani, R. Xu, C. Ventre, B. Horvath, R. Hu, T. Balch, F. Toni, S. T. Kumar (Eds.), Proceedings of the 3rd ACM International Conference on AI in Finance (ICAIF'22), Association for Computing Machinery, New York, NY, 2022, pp. 36–43. doi:10.1145/3533271.3561722.

[7] Deep structural causal models for tractable counterfactual inference, 2020.

[8] B. Barr, M. R. Harrington, S. Sharpe, C. B. Bruss, Counterfactual explanations via latent space projection and interpolation, 2021. URL: https://arxiv.org/abs/2112.00890. arXiv:2112.00890.

[9] M. Pawelczyk, K. Broelemann, G. Kasneci, Learning model-agnostic counterfactual explanations for tabular data, in: Proceedings of The Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 3126–3132. URL: https://doi.org/10.1145/3366423.3380087. doi:10.1145/3366423.3380087.

[10] R. Balasubramanian, S. Sharpe, B. Barr, J. Wittenbach, C. B. Bruss, Latent-cf: A simple baseline for reverse counterfactual explanations, 2021. URL: https://arxiv.org/abs/2012.09301. arXiv:2012.09301.

[11] X. Xiang, A. Lenskiy, Realistic counterfactual explanations with learned relations, 2022. URL: https://arxiv.org/abs/2202.07356. arXiv:2202.07356.

[12] Y. Kawachi, Y. Koizumi, N. Harada, Complementary set variational autoencoder for supervised anomaly detection, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 2366–2370.